DOCUMENT RESUME

ED 076 709 TM 002 718

AUTHOR	Frederiksen, Norman; And Others
TITLE	Development of Provisional Criteria for the Study of Scientific Creativity.
INSTITUTION	Educational Testing Service, Princeton, N.J.
SPONS AGENCY	Graduate Record Examinations Board, Princeton, N.J.
REPORT NO	ETS-RM-3
PUB DATE	Feb 73
NOTE	10p.; Paper presented at annual meeting of American Educational Research Association (New Orleans, Louisiana, February 25-March 1, 1973)
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	College Students; *Creative Thinking; Creativity; *Creativity Tests; *Research Criteria; *Research Skills; Scientific Concepts; Scientific Research; Technical Reports
IDENTIFIERS	*Hypothesis Formulation

ABSTRACT

A test of one aspect of scientific creativity, the ability to formulate hypotheses to account for research findings, was given to 400 college students, along with ability and personality measures. Scores for quantity and quality of hypotheses were reliable and showed evidence of construct validity. Both quantity and quality feedback had their major effect on the quantity of ideas. Development of measures of other aspects of the research enterprise is underway, intended to lead to a set of criterion measures to be used in basic studies of scientific creativity and potentially in the selection and training of creative scientists. (Author)

FORM 6510

PRINTED IN U.S.A.

FILMED FROM BEST AVAILABLE COPY

ED 076709

RM-73-3

RESEARCH MEMORANDUM

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

TM 002 218

DEVELOPMENT OF PROVISIONAL CRITERIA FOR THE STUDY OF SCIENTIFIC CREATIVITY

Norman Frederiksen
Franklin R. Evans and William C. Ward

This paper was presented at the Annual Meeting of the American Educational Research Association, New Orleans, February 26, 1973.

Educational Testing Service
Princeton, New Jersey
February 1973

DEVELOPMENT OF PROVISIONAL CRITERIA FOR THE STUDY OF SCIENTIFIC CREATIVITY*

Norman Frederiksen,
Franklin R. Evans, and William C. Ward
Educational Testing Service

Research in the area of creativity has been handicapped by the lack of reliable and valid criterion measures. Much of the work on creativity has employed tests developed by Guilford and his collaborators to measure divergent production--tests such as Consequences and Brick Uses. Often these tests have been used as criterion measures. The use of a test which was developed to represent one cell in the structure of intellect--the divergent production of semantic units--as a dependent variable interpreted as creativity is not consistent with Guilford's intention and could produce misleading results. Not that it is never appropriate to use Consequences as a dependent measure; it would be highly appropriate to do so when one is trying to increase ideational fluency by a training procedure. But creative performance involves behaviors that are much more complex than ideational fluency, and the two things should not be confused.

A quite different approach to the study of creativity was that of MacKinnon and his collaborators, who studied the characteristics of persons of acknowledged creativity in comparison with those in the same field with lesser degrees of creativity. The method was to secure nominations by members of a profession (such as architects) of those who were most creative, and to invite these and other architects to Berkeley for a

---

*Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, February 26, 1973. The research is being supported by the Graduate Record Examination Board.

period of assessment. Differences found between the most creative people and those of lesser eminence were interpreted as characteristics of creative individuals. The main trouble with this approach is that the characteristics required to attain recognition by one's peers are not necessarily the same as are required to do creative work. In order to be eminent, it may be necessary for the scientist to be a good promoter as well as a good scientist.

The approach we are using is to develop tests of creative performance that may be characterized as intermediate or provisional criteria, using simulations or job samples of the work of a research psychologist (since that is the area we know most about). Such criterion tests, we hope, will elicit interestingly complex behaviors that have a high degree of face validity. If the reliability and validity of such tests can be demonstrated, they should be useful in investigating correlates of various aspects of creative behavior and serving as dependent variables in experiments having to do with situational variables that might influence creativity. It seems feasible to focus the tests and experiments at the level of a second-year graduate student, a time when the student has mastered a significant amount of information but he has not yet become too specialized, and he still may be available (for a fee) to serve as a subject.

In this paper we propose, first, to describe the one such measure for which we have accumulated a fair amount of information about validity and reliability; and second, to describe more generally the battery of tests that is being developed and how they will be used in research on scientific creativity.

The Formulating Hypotheses (FH) test was designed to measure one aspect of scientific creativity: the interpretation of data, i.e., the ability to conceive of hypotheses that might account for research findings. Each item of FH consists of a graph or table showing findings from a research study. The sample item we have been using is a graph showing yearly rates of death from infectious diseases and rates of death from diseases of old age. The finding is stated as follows: "Rate of death from infectious diseases has decreased markedly since 1900, while rate of death from diseases of old age has increased." S is instructed to write hypotheses (possible explanations) that might account for, or help to account for, the finding.

The form of the test used in a recently-completed study* included seven items. They dealt, for example, with data showing that time lost from strikes was greatest in the summer months, and that World War II Navy recruits tested in June and July earned higher test scores on aptitude tests than those tested in other months. Ten minutes per item were allowed.

Five scores were obtained: (1) Number of Hypotheses proposed, (2) Number of Acceptable Hypotheses (a subset of 1), (3) the Average Judged Quality of the hypotheses, (4) the Average Scale Value of the hypotheses (another quality measure based on a scoring method that minimized the influence of such factors as handwriting and quality of writing), and (5) the Average Number of Words per response.

---

*Frederiksen, N., & Evans, F. R. Effects of models of creative performance on ability to formulate hypotheses. Research Bulletin 72-54. Princeton, N. J.: Educational Testing Service, 1972.

Subjects were about 400 undergraduate students at two eastern colleges. They were first given tests of vocabulary and ideational fluency. Then the FH items were administered. A random third of the subjects received feedback after each FH item in the form of lists of model hypotheses which were of high quality. A second third were given models illustrating quantity rather than quality of hypotheses. The remaining subjects were given no models.

The median correlation between items for the Number of Hypotheses score was .39, and the reliability for the five-item posttest was .80. Reliabilities for the other four scores were as follows: Number of Acceptable Hypotheses, .67; Average Judged Quality, .60; Average Scale Value, .48; and Average Number of Words per Hypotheses, .87. Thus, scores of satisfactory reliability can be generated from a free-response test like FH--scores that reflect quality as well as quantity of performance.

Evidence for the independence of these scores and for their construct validity was sought through Multivariate Analysis of Variance. The MANOVAs showed highly significant ($p < .001$) relationships involving both fluency and vocabulary. Ideational fluency was related to the quantity of hypotheses produced (Number of Hypotheses and Number of Acceptable Hypotheses), while vocabulary was related to quality scores (Average Judged Quality, Average Scale Value, and Number of Acceptable Hypotheses). Moreover, feedback treatments produced significant effects ($p < .001$). The quantity model led to higher scores on Number of Hypotheses and Number of Acceptable Hypotheses, and a lower Number of Words per item, while the quality

model resulted in higher scores on Average Judged Quality. These results indicate that meaningful and discriminable indices of quality and quantity of ideas can be derived from the FH test.

The Formulating Hypotheses test used in the study just described did not involve data from psychological investigations; it was a more general test intended for undergraduate students. Our next effort was to develop a similar test made up of items based on psychological studies found in the literature. Items were chosen to vary systematically along two dimensions. One of these concerned the degree of rigor exhibited by the study described; some items represented results from controlled experiments, while others concerned uncontrolled field investigations. Second, items were taken from each of three general areas of psychology: (1) personality-social, (2) learning-educational, and (3) experimental-physiological. This design should make it possible to discover to what extent one's ability to suggest explanations for research findings is dependent on the field from which the problem is drawn, or on the match of this field to the individual's area of specialization.

This version of the test has been administered, using item-sampling procedures, to about 80 graduate students in education. Most of the data have not yet been analyzed, but, at least so far as number of hypotheses is concerned, an interesting effect of item type has emerged. Students were able to produce more hypotheses to account for findings from the personal-social and learning-educational areas than for those from experimental-physiological areas, particularly when interpreting

uncontrolled or poorly controlled studies. We won't know, however, until we've tested individuals with other areas of interest, whether this finding represents a main effect for item type (findings in the personality and learning areas being open to a broader variety of interpretation) or one part of an interaction of item type with the subject's area of interest (education students presumably know more about personality and about learning than they do about physiological processes, while another group of students might show a different and equally plausible pattern in relation to their areas of knowledge). Completing this design will of course be of great importance for our further test-development activities; it remains to be seen whether one version of the FH test will be sufficient for measuring students from all areas of psychology, and perhaps from closely related social science fields as well; or whether tests will need to be tailored for individuals of differing specialities within the field of psychology.

Another concern being investigated has to do with the meaning and generality of quality scores on the FH test. The instructions we have been using ask the student to produce as many reasonable hypotheses as he can to account for each finding; it is not surprising that most students seem to interpret this request as emphasizing the quantity, not the quality, of ideas they can produce. We are currently giving one FH test with instructions emphasizing quantity and another with instructions emphasizing quality to the same set of students, in order to discover whether it is possible to get a good measure of each of these attributes from a single testing or whether different instructions are required to get good measures of quality and quantity of ideas.

We are also engaged in the development of a number of new tests, each intended to sample one aspect of the scientist's productive thinking efforts. This developmental effort is a "bootstraps" operation: Obviously the task would be much more simple and straightforward if we had a coherent, believable theory of scientific creativity from which to sample processes for study. Just as obviously, such a theory is something whose development we hope to contribute to, not something sitting on the shelf. What we have to work with, instead, are suggestions as to processes provided by several sources. Various theories concerning steps in the creative process have been surveyed, beginning with the classic four-stage model for creative problem solving suggested by Wallas in 1926: preparation, incubation, illumination, verification. Flanagan's study of scientists' job performance, using the critical incidents technique, has also been helpful. Guilford's structure of intellect, finally, has been a useful heuristic device. Our approach involves using such sources for suggestions as to possible components of creative scientific thinking; developing one or several measures of each hypothetical component; and then discovering empirically whether the suggestion was a good one--that is, whether reliable dimensions of creative performance that are discriminable from (though not necessarily independent of) other such dimensions can be found. The following measures are the ones under consideration for the first round of this developmental effort; for each, we have given here just the name of the test and an excerpt from the instructions.

Notice that we have chosen names that describe the operation required; we are trying to avoid titles that claim more for the test than we can justify.

Measuring Psychological Constructs. "For each construct, list as many different methods as you can think of for eliciting the behavior implied by the construct, so that it can be observed and measured."

Formulating Research Ideas. "You are at a point in your training where you must choose an area of specialization, and you have narrowed your choice down to two. Your advisor has suggested that, in order for you to get a better impression of the nature and variety of research projects you might engage in, you write down as many research ideas as you can think of in each area. Write titles or brief descriptions of as many research projects as you can think of."

Personnel Selection Problem. "Your boss has asked you to make a list of all the personal characteristics you can think of that might be associated with success in doing the work of a plumber."

Analyzing Psychological Constructs. "There are many constructs in psychology that are usually treated as though they are unitary but may on close examination be found to consist of a number of separate and relatively independent parts. For each construct write names (or short descriptions) of all the parts you can think of that might be identified."

Evaluating Hypotheses. "Here is a list of five hypotheses to account for the finding reported. Which one do you consider the best, i.e., most likely to account for the finding? Rank-order the remaining hypotheses."

Evaluating Proposals. "As a class exercise, you have asked each of your students to write a brief description of a proposed experiment of his own design. For each paper, write your suggestions to the student regarding how the design or methodology might be improved."

Ideational Fluency in Psychology. "Write as many words or phrases as you can think of that have been used to describe personality traits."

Scanning Speed. "You are interested in articles dealing with the effects of anxiety on learning motor skills. Scan the following titles and check the articles that seem relevant and that you might want to read."

This list is long, but it is only a beginning; some of these measures will undoubtedly fall by the wayside, while the need for others will become clear. The hope is that repeated cycles of brainstorming and of empirical tryouts will lead to a relatively compact set of measures, each somewhat distinct from the remainder in the aspect of scientific thinking it requires, and the whole providing a representative sample of those thinking processes psychologists and others must go through in their productive thinking efforts. We may, at the end, be uncertain which of these processes deserve the label "creative"; we may even scrap "creativity" in favor of "productive scientific thinking." In any case, such a battery should provide a vehicle to help in bridging the gap between the simple cognitive abilities, such as those appearing in the structure of intellect, which we may understand but whose usefulness in the world is unclear; and complex cognitive performances whose importance is clear but whose interpretation is beyond us for the present.